EXHIBIT 2

Report on Probabilistic Analysis of Overlaps in Young Adult Paranormal Romance Fiction

Patrick Juola, Ph.D.
Director of Research, Juola & Associates

May 12, 2023

Summary

1 As an expert in computational, quantitative, and forensic linguistics, I have performed a computational and statistical analysis of a 2011 unpublished manuscript of a novel entitled *Blue Moon Rising* compared with the *Crave* series of novels (*Crave, Crush, Covet,* and *Court*). I have been asked to determine the degree of similarity between these works, and in particular, the likelihood that the similarities found could be explained by chance overlap in independent writing, or whether they represent some form of common authorship. As described below, it is my opinion that it is more likely than not that these works share a common origin.

Qualifications

2 I am a Professor of Computer Science at Duquesne University, Pittsburgh, PA, where I hold the Joseph A. Lauritis, C.S.Sp. Endowed Chair in Teaching and Technology. I am also Director of the Evaluating Variations in Language Laboratory at Duquesne.

3 I received a Ph.D. and M.S. in computer science at the University of Colorado at Boulder and an M.S.-level "certificate" in cognitive science from the University of Colorado at Boulder, as well as a B.S. in electrical engineering at the Johns Hopkins University, Baltimore. I received post-doctoral training in experimental psychology (specifically in psycholinguistics) as a postdoc at the University of Oxford, UK. I have published numerous journal articles and book chapters on the computational inference of document authorship via the statistical analysis of linguistic features, and have lectured worldwide on this and related subjects.

4 I am the author of one of the leading works on the subject of authorship analysis and stylometry.

5 I am a frequent ad-hoc reviewer on subjects pertaining to authorship attribution, stylometry, digital humanities, and text analysis for a number of journals, including LLC (formerly Literary and Linguistic Computing), JASIST (Journal of the American Society for Information Systems Technology), and SPE (Software Practices and Experiments). I am a member of the International Association of Forensic Linguistics, and serve as secretary to the American Academy of Forensic Sciences Standards Board (AAFS ASB), specifically to the Questioned Documents Consensus Body.

6 I am the primary architect and designer of the JGAAP (Java Graphical Authorship Attribution Program) authorship analysis system. This system, NSF-funded for nearly $2 million, is one of the best-known and most widely used systems for the analysis of authorship in the world. I have also received nearly $1 million in grants from the Defense Advanced Research Projects Agency (DARPA) for the development of computer security technology based on my work in authorship attribution.

7 I have qualified as a testifying expert in forensic linguistics in the Elizabeth, New Jersey branch of the Executive Office for Immigration Review and in the Court of Protection, Birmingham County Court, Courts of England and Wales (UK). In the USA, I have qualified as a testifying expert in California in Kenney v. Saribalis (FL 1603199, Marin County Superior Court, California, 2018), in Federal court in Chevron v. Donziger, et al., (11 Civ. 0691 LAK, S.D.N.Y.) and in international bilateral arbitration in Chevron Corporation (U.S.A.) and Texaco Petroleum Company (U.S.A.) v. The Republic of Ecuador, PCA Case No 2009-23 (UNCITRAL). I have been deposed

(but was not called upon to testify) in Theranos, Inc. v. Fuisz Pharma LLC (11-cv-05236-YGR, N.D. Calif.) These are all the cases at which I have testified as an expert in trial or deposition. In addition, I am active in the historical and journalistic studies of authorship, most notably in the 2013 unmasking of J.K. Rowling's authorship (under the pen name Robert Galbraith) of *A Cuckoo's Calling.*

8 I am also Director of Research for J Computing, Inc., (dba Juola & Associates, henceforth "J&A") a Pennsylvania corporation specializing in text and authorship analysis. Pursuant to that job, I have been asked to analyze a large set of documents as detailed in the following sections. J&A is being compensated at its usual rate of $500/hr plus expenses for the analysis; I personally am receiving no compensation other than my normal agreement with J&A. Neither the company's compensation nor mine depend upon the outcome of this matter.

Background and Assignment

9 On or about December 3, 2022, J&A was approached by Trent Baer, to evaluate the authorship of two novels and to perform a quantitative analysis to determine the likelihood that the books shared common authorship and that, in particular, that the defendant author had based the wildly successful *Crave* series on the plaintiff author's unpublished manuscript.

Theoretical background

10 There is a well-established academic discipline, called "forensic linguistics,"[1] that specializes in the relationship between language and the law. Among the questions addressed by this field is that of "authorship attribution," the task of determining the author of a questioned document by examining the language used in its writing. Authorship attribution, sometimes called "stylometry" or "stylometrics" is also a well-established (sub)field[2], with applications not only in forensic linguistics, but also in journalism, history, sociology, and many other areas.

11 The basic theory of traditional stylistics and authorship attribution is fairly simple. As McMenamin describes it,

> At any given moment, a writer picks and chooses just those elements of language that will best communicate what he/she wants to say. The writer's 'choice' of available alternate forms is often determined by external conditions and then becomes the unconscious result of habitually using one form instead of another. Individuality in writing style results from a given writer's own unique set of habitual linguistic choices.[3]

12 Coulthard's description is also apt:

> The underlying linguistic theory is that all speaker/writers of a given language have their own personal form of that language, technically labeled an idiolect. A speaker/writer's idiolect will manifest itself in distinctive and cumulatively unique rule-governed choices

[1] McMenamin, Gerald R. *Forensic linguistics: Advances in forensic stylistics.* CRC Press, 2002; Coulthard, Malcolm, and Alison Johnson, eds. *The Routledge handbook of forensic linguistics.* London: Routledge, 2010; Perkins, Ria C., and Timothy D. Grant. "Forensic linguistics." In *Encyclopedia of Forensic Sciences*: Second Edition, pp. 174-177. Elsevier, 2013; Coulthard, Malcolm, Alison Johnson, and David Wright. *An introduction to forensic linguistics: Language in evidence.* Routledge, 2016; Grant, Tim. *The Idea of Progress in Forensic authorship analysis.* Cambridge University Press, 2022. Tayebi, Tahmineh, and Malcolm Coulthard. "New Trends in Forensic Linguistics." Language and Law/Linguagem e Direito 9, no. 1 (2022)

[2] Holmes, David I. "Authorship attribution." *Computers and the Humanities* 28, no. 2 (1994): 87-106; Juola, Patrick. "Authorship attribution." *Foundations and Trends in Information Retrieval* 1, no. 3 (2008): 233-334; Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. "Computational methods in authorship attribution." *Journal of the American Society for Information Science and Technology* 60, no. 1 (2009): 9-26; Stamatatos, Efstathios. "A survey of modern authorship attribution methods." *Journal of the American Society for Information Science and Technology* 60, no. 3 (2009): 538-556; Ainsworth, Janet, and Patrick Juola. "Who wrote this: Modern forensic authorship analysis as a model for valid forensic science." *Wash. UL Rev.* 96 (2018): 1159.

[3] G. McMenamin. Declaration of Gerald McMenamin in *Ceglia v. Zuckerberg and Facebook,* (WD 2012 WL 1392965 (W.D.N.Y)). Available online at https://docs.justia.com/cases/federal/district-courts/new-york/nywdce/1:2010cv00569/79861/50,2011

> for encoding meaning linguistically in the written and spoken communications they produce. For example, in the case of vocabulary, every speaker/writer has a very large learned and stored set of words built up over many years. Such sets may differ slightly or considerably from the word sets that all other speaker/writers have similarly built up, in terms both of stored individual items in their passive vocabulary and, more importantly, in terms of their preferences for selecting and then combining these individual items in the production of texts.[4]

13 These choices express themselves in a number of ways. An easy and obvious example is the use of the written word "honor" instead of the more typical Commonwealth spelling of "honour." A less obvious example[5] is the use of specific function words such as prepositions — for example, is the fork "to the left," "at the left," or "on the left" of the plate in a typical table setting? These differences can be spotted, tabulated, and used to test hypotheses about authorship.

14 In particular, it is highly unlikely that two people would make exactly the same choices when describing the same situation. In recent scholarship[6] I have shown that, even in an extremely constrained domain, such as instructions for making lemonade or a PB&J, people will use different words in their descriptions. (For example, do you use "pieces" or "slices" of bread, and does the J stand for "jam" or "jelly"?) Coulthard[7] has noted that "the occurrence of long identical sequences in two texts is likely to be a product of borrowing"[8], and that a sequence of only seven English words is unlikely enough to be a unique phrase, even against the entire contents of the Google search engine database or Google Books Ngram database. Similarly, Johnson and Woolls[9] write that "most sequences of words are unlikely to be selected and arranged in the same order by two individuals, whether writing on the same topic or not" and that "extended common sequences are even more indicative of a common source." Unique or not, these overlaps are rare enough to provide evidence of common authorship.

The Corpora Analyzed

15 We received various documents for analysis, including electronic copies of the first four books of the *Crave* series (*Crave*, *Crush*, *Covet*, and *Court*, collectively referred hereafter as "Crave") with Tracy Wolff as author of record; electronic copies of manuscripts at various stages of *Blue Moon Rising*, an unpublished novel by Lynne Freeman (the title was later changed to *Masqued*); and ten novels by ten separate authors in the same genre (young adult supernatural/vampire romance, most notably typified by the *Twilight* books/movies) as the disputed novels in question. These additional novels provided an objective baseline to let us control for the specific characteristics of this genre. For *Blue Moon Rising*, to avoid oversampling nonindependent texts (different versions of the same work), we confined our analysis to the 2011 manuscript ("BMR"). All documents analyzed were in English.

16 The ten baseline novels analyzed were:

- *Hush Hush*, by Becca Fitzpatrick
- *City of Bones*, by Cassandra Clare
- *Vampire Diaries: the Awakening*, by L.J. Smith
- *Shiver*, by Maggie Stiefvater

[4] M. Coulthard. *On admissible linguistic evidence.* Journal of Law and Policy, XXI(2):441–466, 2013.

[5] See J. F. Burrows. " 'an ocean where each kind. . . ' : Statistical analysis and some major determinants of literary style." *Computers and the Humanities*, 23(4-5):309–21, 1989; J. N. G. Binongo. "Who wrote the 15th book of Oz? an application of multivariate analysis to authorship attribution." *Chance*, 16(2):9–17, 2003.

[6] Juola, P. "How much overlap means plagiarism? A controlled test corpus." 8th European Conference on Academic Integrity and Plagiarism (ECAIP) 2022, 12, 13–14

[7] M. Coulthard. "Authorship Identification, Idiolect, and Linguistic Uniqueness." *Applied Linguistics*, 25.4-431–447, 2004; See also M. Coulthard, A. Johnson, and D. Wright. Forensic Linguistics: Language in Evidence, 2nd edition. London:Routledge, 2017.

[8] Coulthard, em et al., 2017, p. 189

[9] Johnson, Alison and David Woolls. 2009. "Who wrote this? The linguist as detective." In Susan Hunston and David Oakey (eds.) *Introducing Applied Linguistics: Concepts and Skills.* London: Routledge, 111–118.

- *Twilight*, by Stephanie Meyers
- *Vampire Academy*, by Richelle Mead
- *Wings*, by April Lynne Pike
- *Beautiful Creatures*, by Kami Garcia and Margaret Stohl
- *Fallen*, by Lauren Kate, and
- *The Immortals: Evermore*, by Alyson Nöel

Analysis

17 We first inspected the text of of both Crave and BMR to see if there were any word-for-word overlaps of seven words or more that did not admit of an obvious explanation. Per (Coulthard, 2004, 2007), finding such passages would be strong evidence that either Crave had quoted BMR, that BMR had quoted Crave, or that both authors had independently quoted a common source.

18 As described above, Coulthard (2004; 2017) has argued that any string of seven words that does not contain a common stereotyped phrase is extremely rare and thus strong evidence of shared authorship.. As a simple example, the phrase "my sister's orange cat likes to eat" which seems innocuous enough and which contains seven words does not appear in the Google search engine as a phrase.[10]. In fact, the phrase "orange cat likes to eat," a five word phrase, only appears a few times in the Google search engine and does not appear in the Google Books Ngram viewer[11]. It should be noted that seven words is not a hard cutoff; a rare phrase can be a shorter phrase such"trombone playing scuba instructor," which is not by itself objectionable but is nevertheless rare enough not to appear in Google as a phrase[12].

19 We did find several such phrases shared between Crave and BMR, of which "in the air as I try to", "that million-dollar smile of his", and "on my arms and the back of my neck" are the most notable. (Others, such as "the ground to open up and swallow me" can be easily explained as a common idiom/expression.) The phrase "on my arms and the back of my neck" appears in roughly 20 independent contexts in Google search[13] which is rare but still present. Similarly, "that million-dollar smile of his" is only six words long but occurs only 3,370 times among the hundreds of billions of web pages indexed by the Google search engine. The phrase "in the air as I try to" appears only seven times in this set.

20 Thus, "in the air as I try to" qualifies as a seven-word lexical overlap and is, in Couthard's view, unlikely to be independently written by two different authors. The others are five- and six-word overlaps that further confirm this improbability even if they do not Using Couthard's methodology, we therefore conclude that it is more likely than not that these phrases come from a common source and that that source is unlikely to be a publicly accessible document. However, this is not a quantitative assessment of probability. To get this, we need to use mathematics.

21 We therefore took advantage of Baer's prior work in analyzing the relevant books. He sent us a list of twenty-five commonalities among various documents that he considered noteworthy, of which six related specifically to the 2011 version of BMR that we analyzed. Of these commonalities, we independently reanalyzed them to determine, first, if we could confirm their presence in BMR *and* in at least one of the Crave books; and secondly; whether or not they appeared in the Google Books Ngram Database[14]. This database provides access to the various word clusters that are found in the books collected by the Google Books project.

[10]Phrasal search for "my sister's cat likes to eat" (including quotation marks). Google.com, accessed May 8, 2023

[11]Search for "orange cat likes to eat" https://books.google.com/ngrams. Accessed May 8, 2023

[12]Phrasal search for "trombone playing scuba instructor" https://google.com. Accessed May 11, 2023

[13]Phrasal search for "on my arms and the back of my neck" (including quotation marks). Google.com, accessed May 8, 2023.

[14]https://books.google.com/ngram; see also Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* (Published online ahead of print: 12/16/2010)

22 Of the six commonalities related to the 2011 version identified by Baer, we confirmed four to be of further interest as not occurring in the Google Books Ngrams viewers. These commonalities were:

- "tree stump seats" in BMR; a similar phrase appears as "tree-stump seats" in the Crave series.
- "small Stonehenge" appears in BMR; "Stonehenge lite" appears in Crave.
- "air as I try to" appears in both, but in wildly different contexts:
  - my arms flailing wildly *in the air as I try to* catch my balance. (BMR)
  - I'm gasping for *air as I try to* make it through the roiling clouds. (Crave)
  - The words hang *in the air as I try to* absorb them (Crave)
- "that million-dollar smile of his" (note hyphen) appears in both
  - he just grinned *that million-dollar smile of his.* (BMR)
  - the corners of *that million-dollar smile of his* are wilted (Crave)

23 Note that the the last two items in the list immediately preceding are also examples of long lexical overlaps.

24 We will consider the phrase "air as I try to," a five-word subsequence of item 3 above. We will now proceed to quantify the probability of this phrase appearing by chance in two different and independently-written works.

25 Google Books N-grams (henceforth GBN) contains the contents of roughly 15,000,000 books. These are presented as clusters ranging from isolated single words to five adjacent words (5-grams). For example, "in the beginning" would be a 3-gram as it contains three words. At 10,000 words per book (remembering that not all "books" in GBN are novel-length), this becomes 150,000,000,000 (150 billion) words in the GBN database. For any common word or phrase, it is extremely likely that it would be found somewhere in this huge sample of English. Conversely, a phrase that does not appear is simply an uncommon phrase.

26 For a phrase to not appear anywhere in a work (such as Crave), this means that it does not occur as the first phrase *and*, independently, it does not appear as the second phrase *and* it does not appear as the third phrase, and so on. (See the appendix for technical details.)

27 As discussed in the appendix, the four books of the *Crave* series contain 868,455 words. This is equivalent to buying about 900,000 tickets in a lottery with an extremely small probability of winning. While your chances of winning are certainly better if you buy lots of tickets, you may still not win. In this instance, we have calculated the probability of "winning," that is, of finding a very rare word or phrase in Crave as only six hundredths of a percent.

28 Continuing our analysis, we calculate the probability of finding a word in BMR that is also in Crave, but that is *not* in GBN to be a tiny fraction of a percent, substantially less than one chance in 1000. Of course, this calculation assumes that Crave and BMR were independently written; if there were shared authorship, the degree of similarity could be much higher.

29 In old-fashioned statistics, the convention is that only results with associated probabilities of 0.05 (5%) or less are "significant" and should lead to a rejection of the hypothesis that these documents were independently written. Modern statistics has rejected this viewpoint[15] and focuses instead on the specific meaning of a p-value—how likely a result is *if* a particular hypothesis is true. In this case, our "particular" hypothesis is that the Crave series and BMR are independently written. If this were true (if the documents had been independently written), this kind of similarly would be seen only once per thousand times.

30 More formally, 999 times out of a thousand we would see no overlap of this type if there were no shared authorial influences, so it is more likely than not that if the manuscripts were composed independently, there would be no overlap. Since we have observed that there IS overlap (and overlap is the rare case), it follows that it is very much more likely than not that they were not independently written.

[15] Ronald L. Wasserstein and Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

31 An identical calculation applies to the phrase that "million-dollar smile of his," which also appears in both Crave and BMR, and which, independently, would only be expected to occur 6% of the time. This provides two different and independent pieces of evidence, either of which suffices to imply that is is more likely than not that these works share authorship. However, taken in conjunction with each other, they provide much stronger evidence for the simple reason that two unlikely events are that much more unlikely than one. To calculate the effects exactly would require some tedious math, but we can approximate this by simply multiplying the probabilities of the two events. One chance in a thousand times one chance in a thousand is roughly once chance in a millions.

32 The calculations described above apply only to exact, word-for-word matches. To the extent that there are other, as yet unknown, word-for-word matches between BMR and the Crave series, the same argument and probabilities would apply, making it that much more "more likely than not" that the writing process was not independent.

33 For inexact matches such as "tree stump seats" / "tree-stump seats" or "small Stonehenge" / "Stonehenge lite," the calculations do not strictly apply and are more complex, but it is intuitively obvious that overlaps of rare concepts like these are notable and unlikely to occur independently in two sources. Google Books identifies the ten most common words that follow "Stonehenge" and none of them are "lite." In fact, none of them are adjectives—they are only common verbs, conjunctions, or prepositions. Similarly, the ten most common words that precede "Stonehenge" are prepositions, articles, or the start of a sentence—not adjectives. Thus we conclude that these rare cooccurrences are further evidence of the unlikelihood of independent writing atop the quantitative evidence above.

34 Finally, my attention has been drawn to the proper names "Katmere" (Crave) and "Katmai" (Freeman copyrighted *The World.doc* note), which both contains the highly unusual character cluster 'katm.' A well-respected website[16] identifies only five words with this cluster, four proper nouns of non-English origin (such as "Katmandu") and one initialism. This cluster does not appear anywhere in the million-word Brown corpus[17] . Thus it appears with frequency less than 1/1,000,000 words.

35 We can carry out probability calculations on this cluster. Given that it appears in the BMR-related document, the probability of it appearing somewhere in Crave is The probability of 'katm' appearing anywhere in the 868455 words of Crave is 0.580. The probability of it not appearing anywhere in the million words of the Brown corpus is roughly 0.368. Thus, the probability of this cluster appearing in both BMR and Crave is 0.213 (21.3%), which is unlikely-but-not-incredible.

36 It finally remains to confirm that the specific features studied are not, while uncommon in general English, common aspects of the genre that both works share. To this end, we collected ten different YA vampire novels by ten different authors and confirmed that neither "tree stump seats," "Stonehenge lite," "air as I try to," nor the 'katm' letter cluster appear anywhere in these ten novels. Thus we conclude that these phrases are not an integral part of the supernatural fiction genre.

37 We thus have four independent analyses showing that, while it is not unbelievable that the Crave series and BMR are independently written, the statistics show that it is more likely than not that they both derive from the same specific sources containing these linguistic features. We have also excluded genre conventions as a possible source for these features.

Scientific Validity of this Type of Analysis

38 While each court and court system of course makes its own decisions about the admissibility of scientific or technical evidence, both the National Research Council[18] and the President's Council of Advisors on Science and Technology[19] have written about the need for scientific validation for forensic science. PCAST, in particular, stresses [p. 4] the need for "scientific standards" for validity, including demonstrating that proposed methods are "*repeatable*, *reproducible*, and *accurate*, at levels that have been measured and are appropriate to the intended application." [emphasis in original]. (This corresponds to the legal requirement, in FRE 702(c), of "reliable principles and methods.").

[16] `https://www.thefreedictionary.com/words-containing-katm`

[17] Ku cera, H., and Francis, W.N. 1967. *Computational Analysis of Present-day American English.* Providence: Brown University Press.

[18] *Strengthening Forensic Science in the United States: A Path Forward*, August 2009

[19] *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods,* September 2016

39 The reliability and hence admissibility of forensic linguistics have been studied extensively in the literature.[20] Numerous studies have confirmed the effectiveness of stylometric authorship attribution. A recent law review article[21] provides numerous examples of court cases resolved in part through the use of stylometric evidence. More relevantly, it provides an argument in favor of the admissibility of this type of evidence. Similarly, the quantitative and probabilistic analysis of language is a well-studied field, with its own journals, research groups, and conferences. The computational analysis of language of course goes back to work in the 1950s such as the Georgetown-IBM experiment in machine translation; the journal MT (Mechanical Translation) began publication in 1954[22].

40 Briefly reviewing the *Daubert* factors[23], I note that (1) this technique has been extensively tested, including the papers cited above; (2) there is an extensive history of peer review and publication dating back to the 1960s[24]; (3) the error rate can be easily derived from the testing regime; (4) standards and protocols have been proposed for these problems[25] although they do not have regulatory force, and (5) a substantial research community focused on stylometry exists[26]. Google Scholar lists more than 8,000 publications including the word "stylometry" in its catalog. This demonstrates that computational stylometry of the sort described in this report is, as PCAST demands, "repeatable, reproducible, and accurate" and therefore can be a reliable basis for an evidentiary opinion.

Conclusions

41 Quantitative and computational analysis indicates that there are similarities between the *Crave* novels and the 2011 manuscript for *Blue Moon Rising*, and that these similarities are unlikely to have arisen by chance. Based on the evidence described in this report, I consider it to be more likely than not that these works have a common origin. In fact, based on the calculations above, I would consider the odds in favor of these works having a common origin as more than 1000:1[27] A similar but independent analysis of orthographic overlap independently shows that Crave and *The World.doc* are more likely than not to share common origins, with the odds of at least 4:1.

42 I reserve the right to revise and to resubmit this report if I receive new evidence, to correct or to clarify any errors or confusion, or for any other lawful purpose.

Respectfully submitted,

Patrick Juola, Ph.D.
Juola & Associates

[20] See Chaski, C. E. "The Keyboard Dilemma and Forensic Authorship Attribution." *Advances in Digital Forensics*, vol. 3, 2007; Chaski, Carole E. "Best practices and admissibility of forensic author identification." *Journal of Law and Policy* 21 (2012): 333; McMenamin, G. "Disputed Authorship in US Law." International Journal of Speech, Language and the Law, vol. 11, no. 1, 2004, pp. 73–82.; Juola, Patrick. "Authorship Attribution." *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, 2007, pp. 233–334.

[21] Ainsworth, Janet, and Patrick Juola. "Who wrote this: Modern forensic authorship analysis as a model for valid forensic science." Wash. UL Rev. 96 (2018): 1159

[22] K. S. Jones, "Natural Language Processing: A Historical Review." *Current Issues in Computational Linguistics: In Honor of Don Walker (Linguistica Computazionale)* (9–10):3–16. 1994

[23] `https://www.law.cornell.edu/wex/daubert_standard`

[24] Mosteller, F. and D.∼L.∼Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers. Addison-Wesley, 1964.*

[25] Chaski, Carole E. "Best practices and admissibility of forensic author identification." *Journal of Law and Policy* 21 (2012): 333; Juola, Patrick. "The Rowling case: A proposed standard analytic protocol for authorship questions." *Digital Scholarship in the Humanities* 30.suppl_1 (2015): i100-i113

[26] See, for example, the Computational Stylistics Group at `https://computationalstylistics.github.io/` or the Digital Literary Stylistics Special Interest Group (SIG-DLS) at `https://dls.hypotheses.org/`

[27] These number come from the analysis of Crave and the 2011 manuscript of *Blue Moon Rising*.

Technical Appendix

43 The calculations described in the main body of this report rely on three basic statistical principles:
First, the total probability of anything that could happen sums up to 1.0 (100%). Secondly, if the probability of an event E happening is X, the probability of E not happening is 1.0-X. (for example, if my team has a 0.65 (65%) chance of winning the game, then the chances of not winning,—whether we lose, tie, or get rained out, etc.—is 1-0.65 or 0.35). Thirdly and finally, if E and F are independent events, with E having a probability of P_E and F having a probability of P_F, the chance of both of these happening are the product of the probabilities P_E and P_F). Again, if the probability of my football team winning is 0.65 and the probability of a thunderstorm is 0.10, the chance of both happening—my football team winning in a thunderstorm—is 0.065.

44 From these two facts, it follows immediately that the chance of my football team not winning is 0.35 and the chance of there not being a thunderstorm is 0.9, and the chance of both not-X and not-Y happening is 0.35 times 0.9 or 0.315.

45 We can use this to calculate the likelihood of a phrase appearing (as in para. 21 *et seq.*, above) as follows:

a) We begin by assuming that English words[28] have a probability distribution and that the words in the Google Books database, Crave, and BMR are all drawn from this distribution. A word or phrase only appears in GBN if it has at least forty occurrences in the 15,000,000 books (roughly 150,000,000,000 words) database of Google Books. We use this threshold to define "rare" words (words that occur in English with probability less than or equal to 40/150,000,000 (∼0.000000000266666666666666... or 2.6666667 x 10^{-10}).

b) The chance of any specific word appearing at any specific place in a randomly chosen text being rare is therefore at most 40/150,000,000,000 (∼0.000000000266666666666666... or 2.6666667 x 10^{-10}). This is equivalent to taking a random book off a library shelf, putting a pin down on a random page, and checking to see if GBN has any hits. Most words or phrases will of course be found, but, rarely, the pin will select something like "Parasaurolophini" or "trombone playing scuba instructor."

c) The chance of selecting a word or phrase that is not rare is 1.0 minus the probability derived from a), or (roughly) 0.99999999926666.

d) The chance of selecting two words and having them both be rare is the probability from a) squared, while the chance of selecting two words and having neither of them being rare is the probability from b) squared.

e) By extension, the probability of selecting a large number of words and having none of them be rare is the probability of a not-rare word [from b)] raised to the power of the number of words. (E.g., 100 not-rare words would be 0.99999999926666^{100}).

f) The probability that at least one word in a large number of words is rare is thus 1.0 - the probability from e).

g) We now consider the probability that a rare word is in the text of Crave. Since the first four books of the Crave series have 868,455 words, the probability of that no words in Crave are rare is $1.0\text{-}0.99999999926666^{868,455}$ (∼0.9994 or 99.04%) and the corresponding probability that at least one word in Crave is rare (because "at least one word" means "no words" didn't happen) is 1.0 - 0.9994 or 0.0006 (0.06%). This number is higher than the probability of any specific word being rare because there are such a huge number of words and phrases in a four book series and thus so many possible chances to find one—this represents the fact that if enough people buy a lottery ticket or the equivalent, someone is eventually likely to win.

[28] Here and after, the word "word" includes short phrases of the sort stored in the Google Books database unless otherwise noted.

h) We are now in a position to estimate the likelihood of a word being a rare word that appears in Crave but does not appear in GBN. The probability of this is the probability of a word not appearing in GBN [from a)] *and* also appearing in Crave [from g)] or 0.000000000266666666666666 times 0.0006, which is 0.000000000000001599999999999 or roughly 1.5 x 10^{-13}.

i) We remind the reader that we are looking for (rare) words/phrases that appear in BMR, appear in Crave, but that do not appear in GBN. In order to be such a word, it needs to fulfill three conditions. First, it must appear in BMR, which any word or phrase chosen from BMR will automatically fulfill. Second, it must not appear in GBN, which we have previously calculated to occur with probability ∼0.0000000002667. Finally, it must (independently) appear at least once in Crave. The product of these probabilities is the same 1.5 x 10^{-13} calculated in the prior paragraph.

j) The previous calculation holds not only for any of the 150,140 individual words in BMR, but also for the 2-, 3-, 4- and 5-grams that begin at each individual word. There are therefore 5 times 150,700 (780,700) opportunities in BMR for such a word/phrase to be found. Again, this is similar to a lottery with nearly 800,000 tickets, but a winning chance measured in fractions of trillionths. As before, probability of someone winning is 1.0 - the probability of no one winning, which in turn is the probability of any specific person raised to the power of the number of opportunities.

k) (1 - 1.5 x $10^{-}13)^{780,700}$ (= ∼0.999999875) is the probability that no one wins the lottery, which in turn makes the probability that there is at least one "winner" substantially less than one chance in 1000.

46 It should be noted that there is substantial margin for error in the various estimates made in this calculation, but that these errors would not affect the ultimate opinion. Even if we assume instead that the threshold of rare words is 100 times greater (2.6666667 x 10^{-8} instead of 2.6666667 x 10^{-10}), the probability of no rare words in Crave is less than 3%, the probability that at least one word in Crave is rare is about 97%, the probability that any given word is in Crave but not in BNR is roughly 6.1 x 10^{-10}, and the chance that one of the words or phrases in BMR is in Crave, but not in GNB, is still less than 1 chance in 1000.

47 Finally, it should be noted that the analysis above makes a rather questionable independence assumption; it assumes that documents are independent of each other (which is legitimate), but also that words in a document are independent of each other, which is sometimes untrue. Documents are "about" something, and seeing one word suggests that other, related, words will appear more often. For example, a novel where a character is a vampire will systematically use the word "vampire" more often than in a non-fiction book about the Vietnam War.

48 However, this only applies to topical elements of the story and genre. None of the features studied in this report appear to be topical. Furthermore, if they were topical elements, it would be expected that these features would show up in other examples of the genre. Since these elements don't show up in the unrelated control samples, I consider the independence assumption to be justified in this instance.